IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Rajesh Murari

August 10, 2024

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

- **In this capstone project, we will predict if the SpaceX Falcon 9 first stage land successfully using several machine learning classification algorithms.**
- The steps involved in this project are:
  - Data Collection, and Data Wrangling
  - Exploratory Data Analysis
  - Interactive Data Visualization
  - Machine Learning Prediction
- Summary of all results
  - Exploratory Data Analysis results show that some features of the rocket launches have a correlation with outcome of launches.
  - Predictive Analysis Results, shows that Decision Tree is the best machine learning algorithm to predict if the Falcon 9 first stage will land successfully.

Introduction

- **Project background and context**
  - In this capstone, we will predict if the Falcon 9 first stage will land successfully. SpaceX advertises Falcon 9 rocket launches on its website, with a cost of 62 million dollars; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.
- **Problems want to find answers**
  - The main question that we are trying to answer is, for a given set of features about a Falcon 9 rocket launch which include its payload mass, orbit type, the location and proximities of a launch site, i.e., the initial position of rocket trajectories, will the first stage of the rocket land successfully?

Section 1

Methodology

Methodology

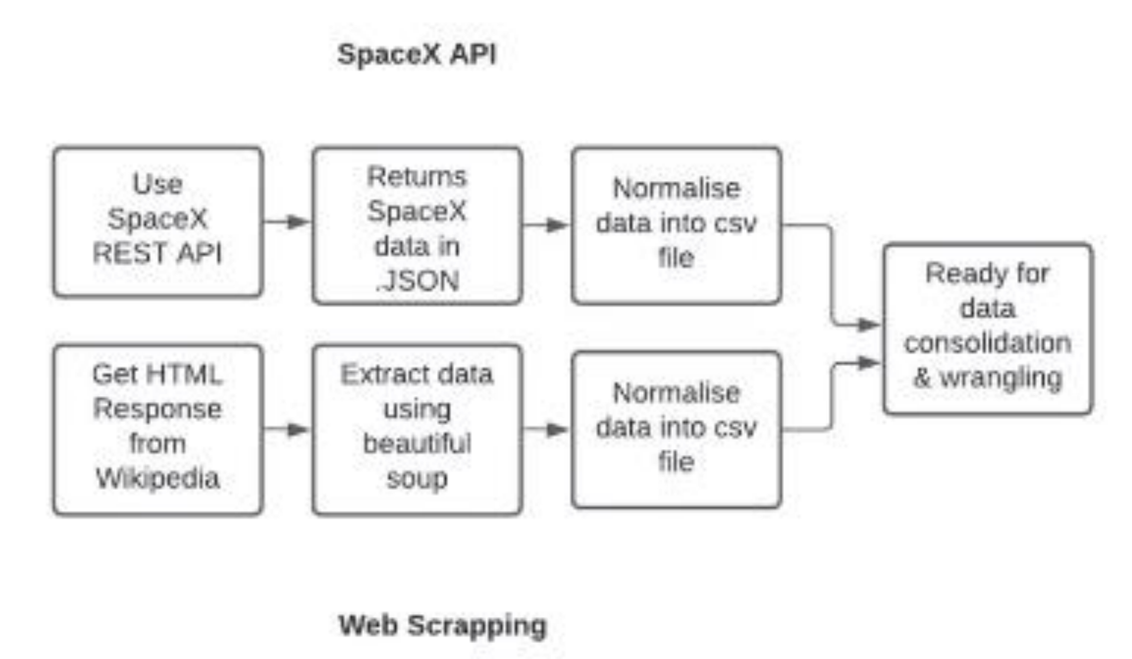
- The overall methodology includes:
  1. Data collection, wrangling, and formatting using::
     - SpaceX Rest API
     - Web Scrapping from Wikipedia
  2. Exploratory Data Analysis (EDA), using:
     - Numpy and Pandas
     - SQL
  3. Data Visualization using:
     - Matplotlib, and Seaborn
     - Folium
     - Dash
  4. Machine Learning Prediction using:
     - Logistic regression
     - Support Vector Machine (SVM)
     - Decision Tree
     - K-nearest Neighbors (KNN)

1. Data Collection, Wrangling, and Formatting

- Data collection process involved a combination of API requests from SpaceX REST API (https://api.spacexdata.com/v4/rockets/) and Web Scraping data from a table in SpaceX's Wikipedia entry (https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922)
- **Data Columns are obtained by using SpaceX REST API:**
  - FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LangingPads, Block, ReusedCount, Serial, Longitude, Latitude
- **Data Columns are obtained by using Wikipedia Web Scraping:**
  - Flight No., Launch Site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booser, Booster landing, Date, Time.

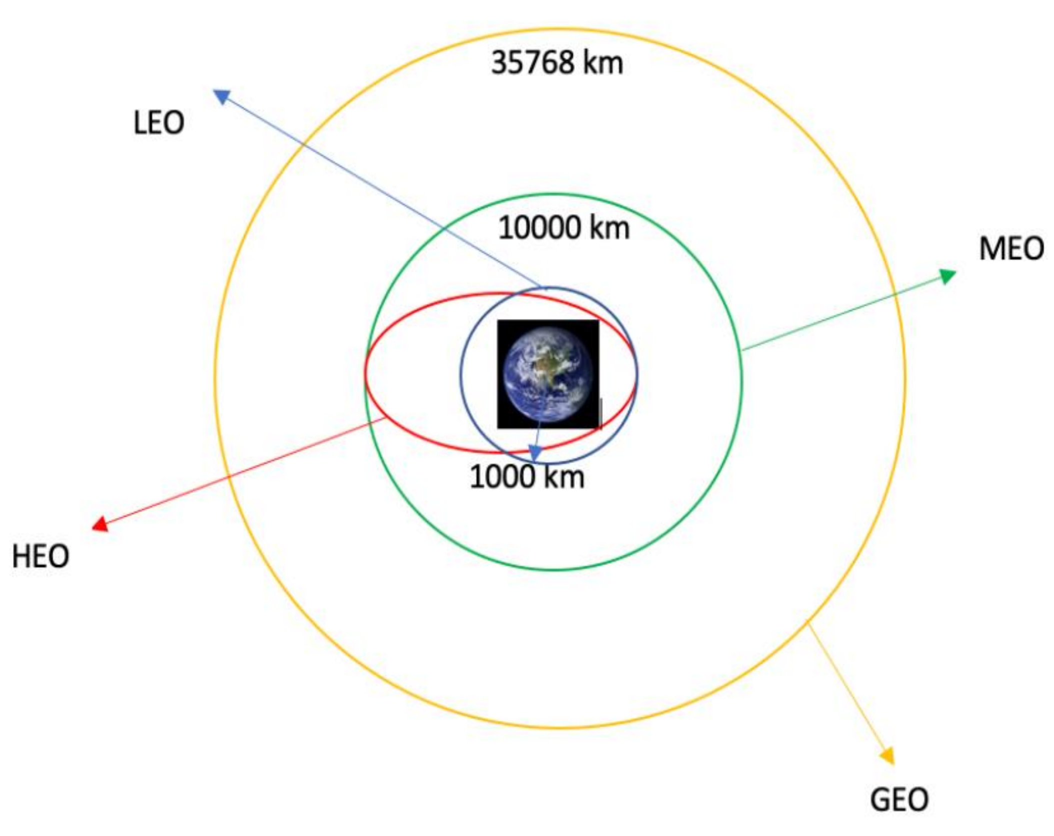
Data Collection and Web Scraping

- Data collection with SpaceX REST calls
- Data Scraping from Wikipedia
- https://github.com/RajeshMurariGitHub/Data-Science-and-Machine-Learning-Capstone-Project/blob/main/jupyter-labs-spacex-data-collection-api.ipynb


SpaceX API
Use SpaceX REST API
Returns SpaceX data in .JSON
Normalise data into csv file
Ready for data consolidation & wrangling
Get HTML Response from Wikipedia
Extract data using beautiful soup
Normalise data into csv file
Web Scrapping

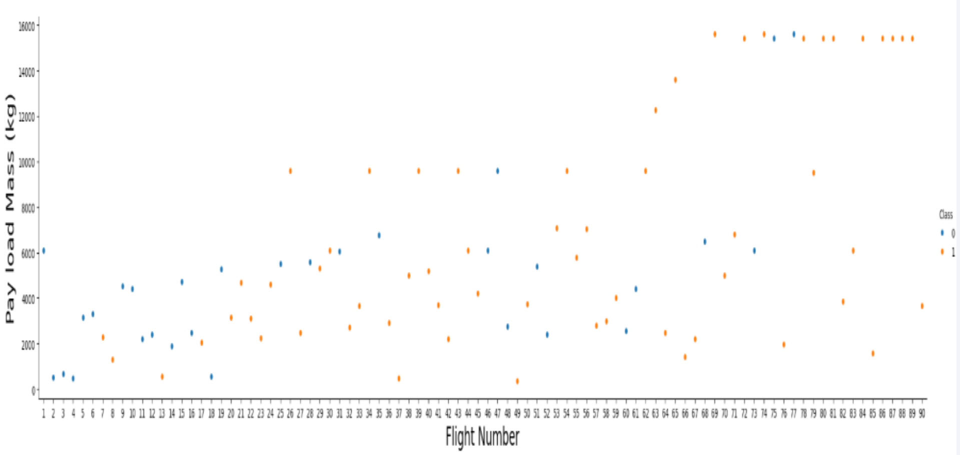
Data Wrangling

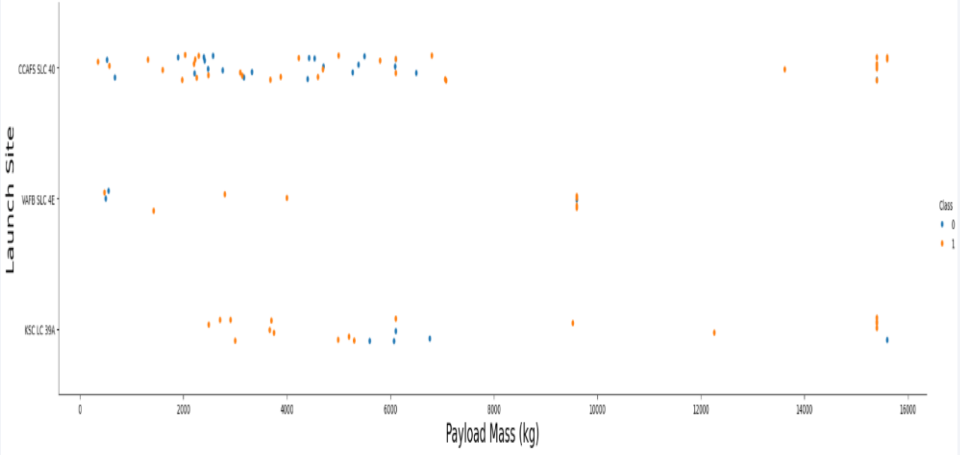
- Data Wrangling:
  - We perform exploratory data analysis and determined the training labels.
  - We calculate the number of launches at each site, and the number and occurrences of each orbit.
  - We created landing outcome label from outcome column and exported the result to csv
  - The link to note book is:
  - https://github.com/RajeshMurariGitHub/Data-Science-and-Machine-Learning-Capstone-Project/blob/main/labs-jupyter-spacex-data_wrangling_jupyterlite.jupyterlite.ipynb



EDA with Data Visualization

- **Pandas and Numpy:**
  - The number of launches on each launch site
  - The number of occurrence of each orbit
  - The number and occurrence of each mission outcome



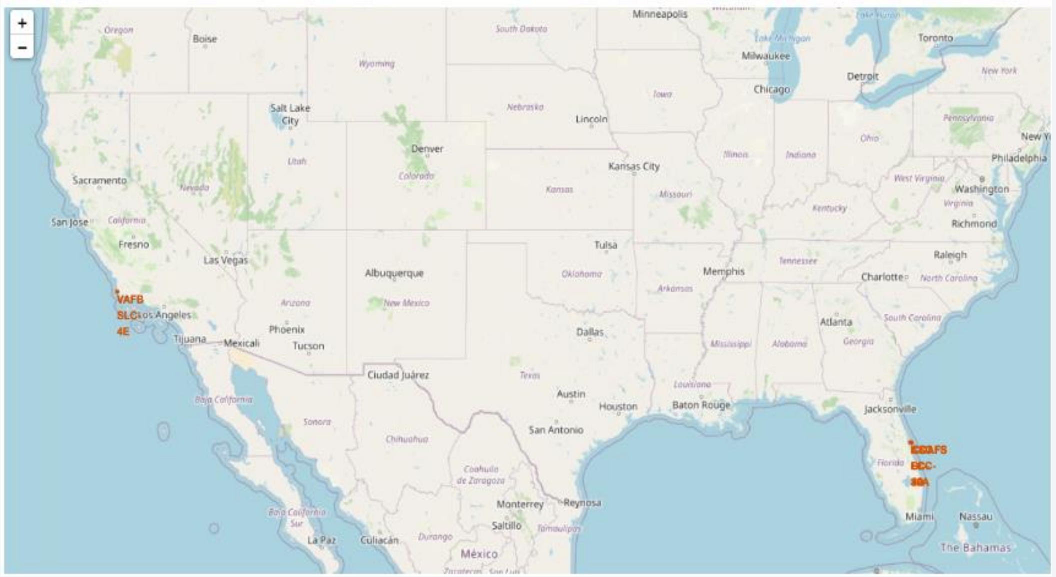


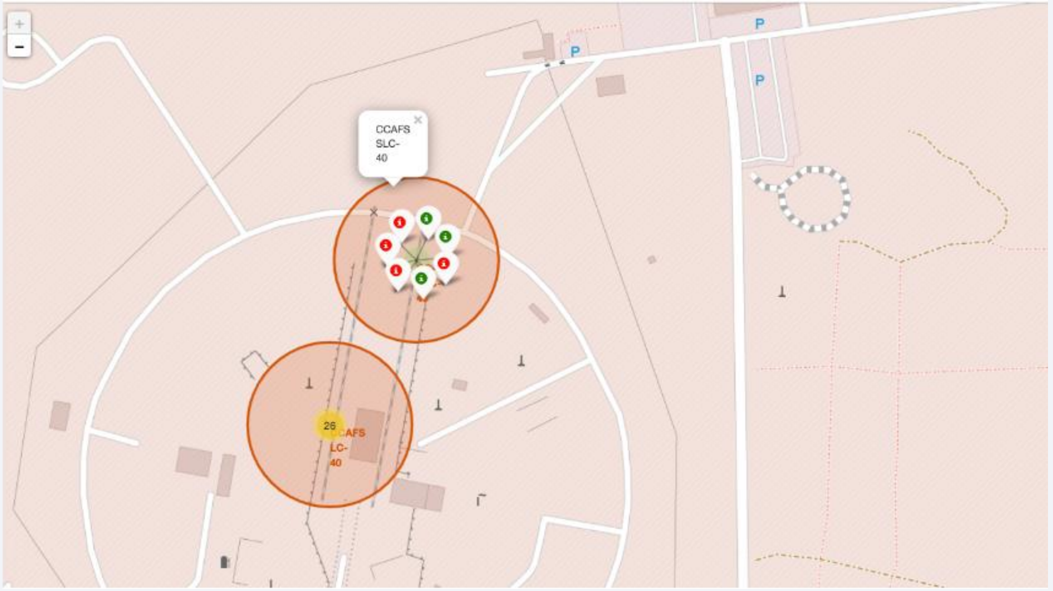
- https://github.com/RajeshMurariGitHub/Data-Science-and-Machine-Learning-Capstone-Project/blob/main/jupyter-labs-eda-dataviz.ipynb.jupyterlite.ipynb

EDA with SQL

- SQL queries performed include:
  - Display the names of the unique launch sites in the space mission
  - Display 5 records where launch sites with the string ‘KSC’
  - Display the total payload mass carried by boosters launched by NASA (CRS)
  - Display average payload mass carried by booster version F9 v1.1
  - List the date where the successful landing outcome in drone ship was achieved.
  - List the names of the boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000.
  - List the total number of successful and failure mission outcomes.
  - List the names of the booster versions which have carried the maximum payload mass. Use a subquery.
  - List the records which will display the month names, succesful landing_outcomes in ground pad ,booster versions, launch_site for the months in year 2017.
  - Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order
- https://github.com/RajeshMurariGitHub/Data-Science-and-Machine-Learning-Capstone-Project/blob/main/jupyter-labs-eda-sql-edx_sqllite.ipynb

Build an Interactive Map with Folium





- All launch sites considered in this project are in very close proximity to the coast While starting rockets towards the ocean we minimize the risk of having any debris dropping or exploding near people.

- From the color-labeled markers in marker clusters, you should be able to easily identify which launch sites have relatively high success rates..

Predictive Analysis (Classification)

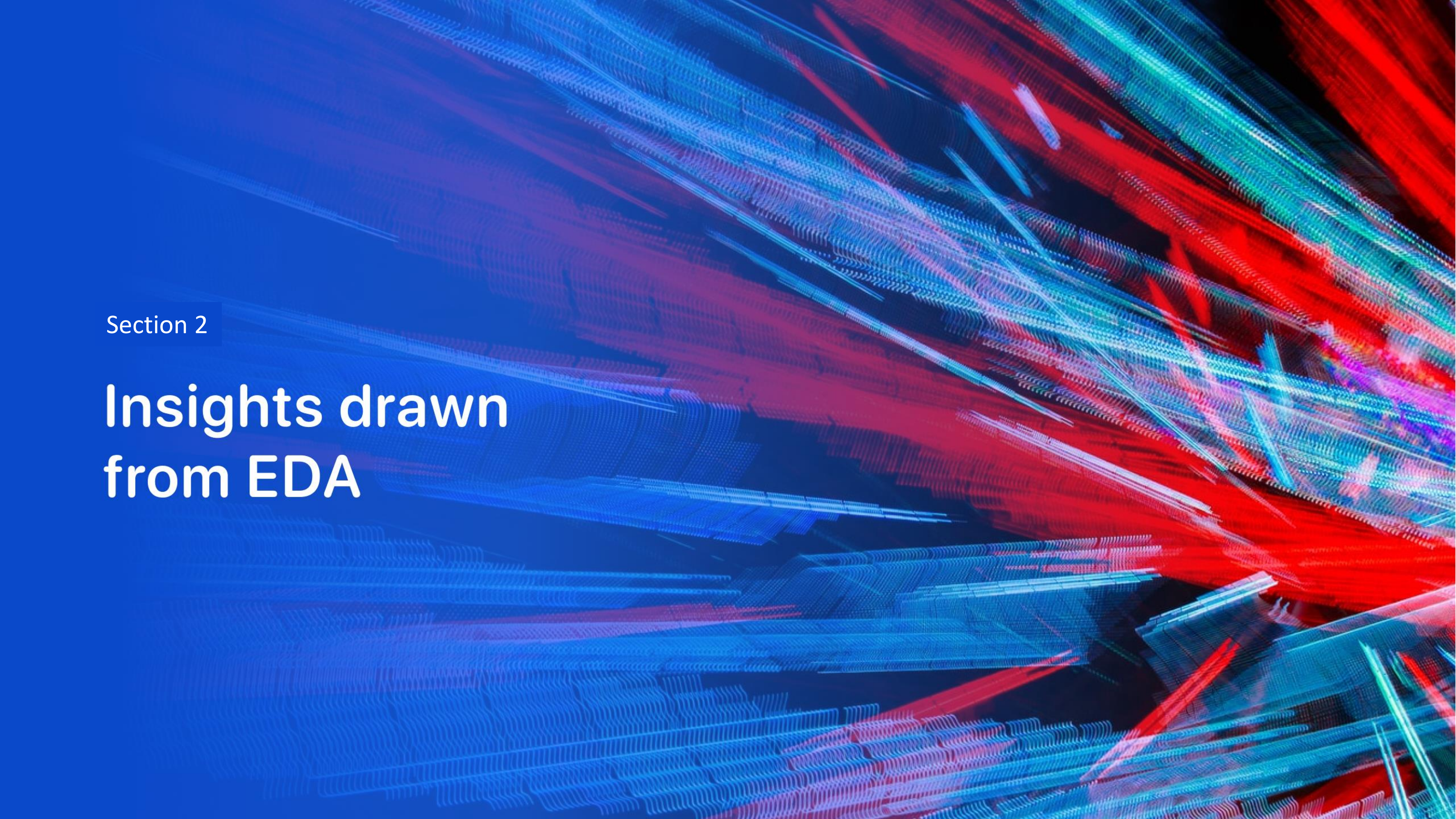
- We loaded the data using numpy and pandas, transformed the data, split our data into training and testing.
- We built different machine learning models and tune different hyperparameters using GridSearchCV.
- We used accuracy as the metric for our model, improved the model using feature engineering and algorithm tuning.
- We found the best performing classification model.
- The link to the notebook is https://github.com/RajeshMurariGitHub/Data-Science-and-Machine-Learning-Capstone-Project/blob/main/SpaceX_Machine_Learning_Prediction_Part_5.jupyterlite.ipynb

Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

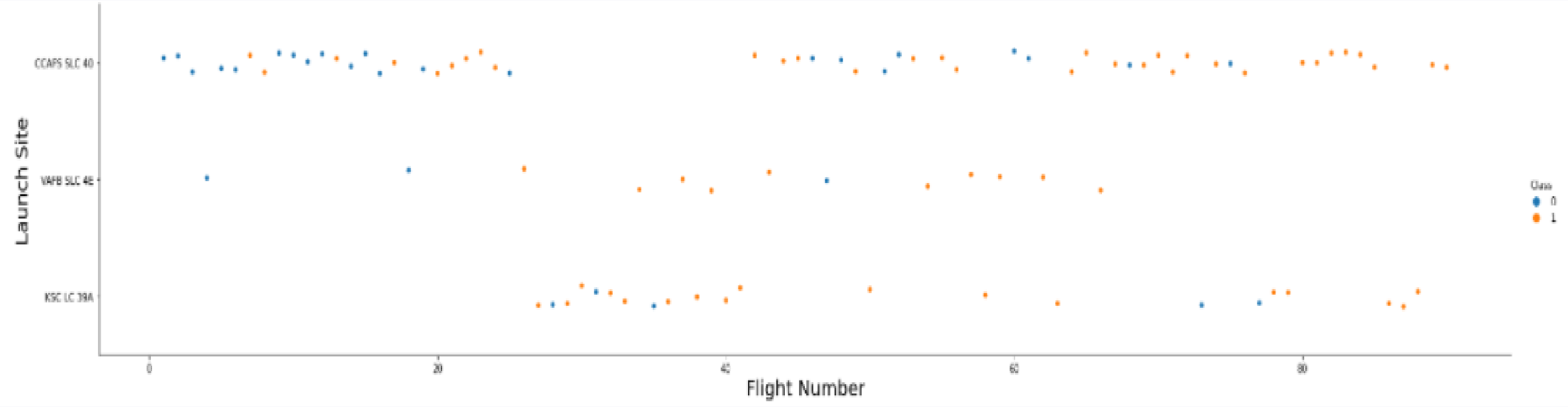
Section 2

Insights drawn from EDA



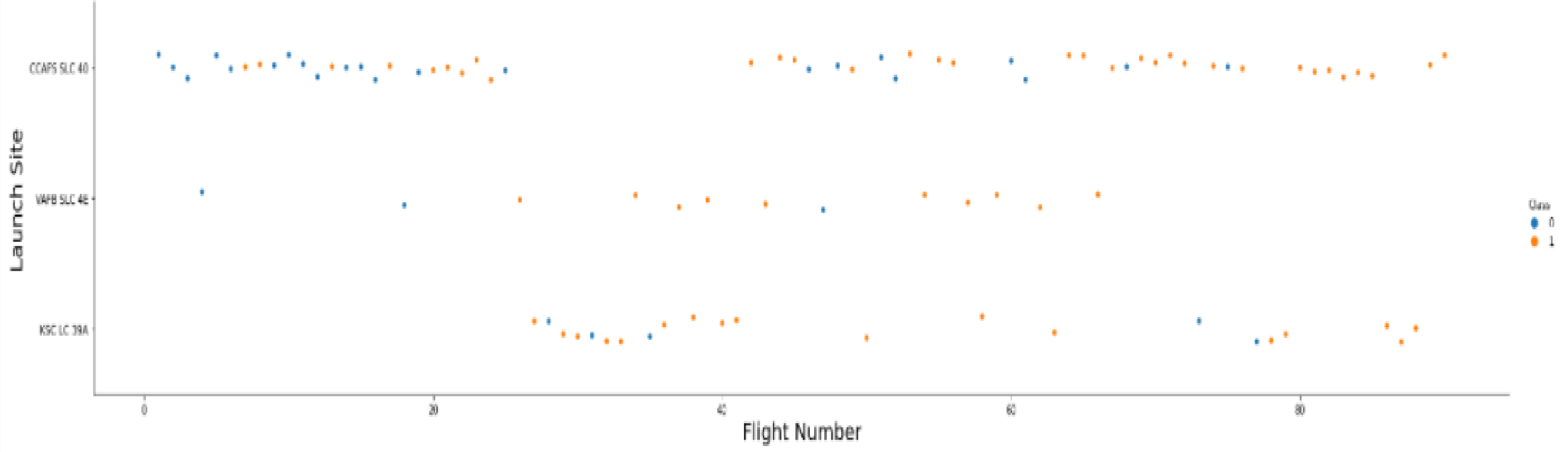
Flight Number vs. Launch Site

- From the plot, we found that the larger the flight amount at a launch site, the greater the success rate at a launch site.



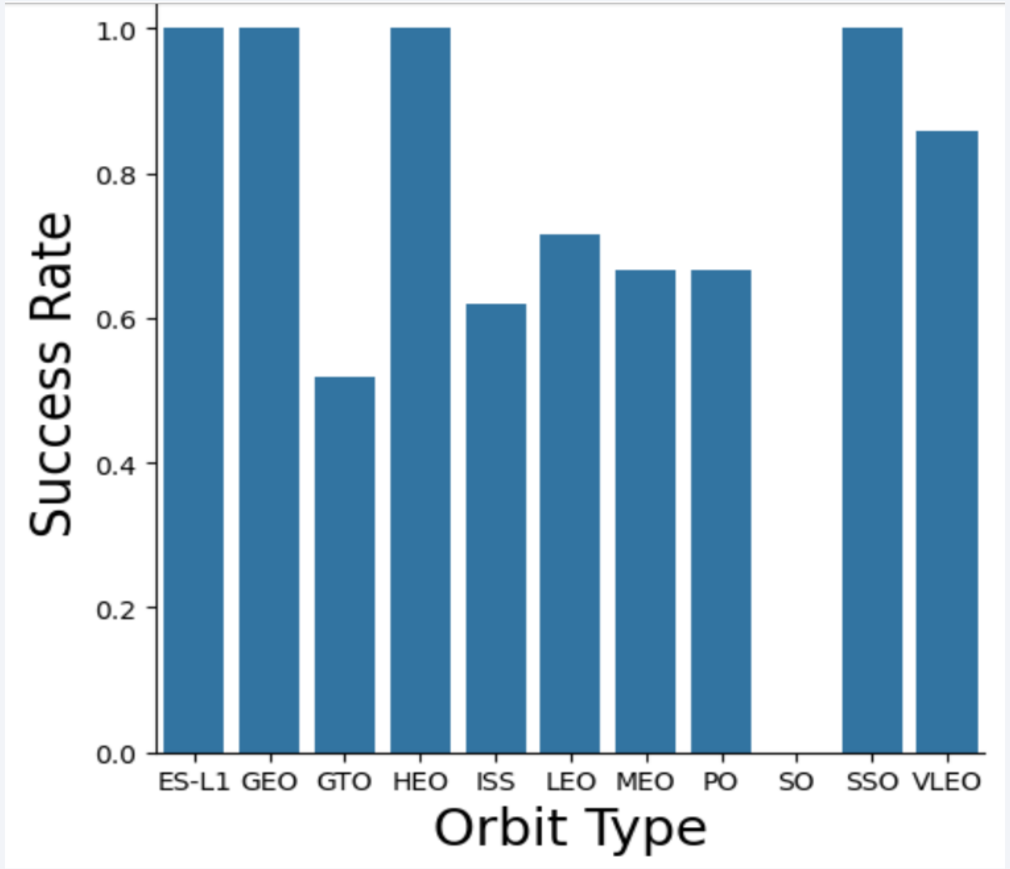
Payload vs. Launch Site

- The greater the payload mass for launch site CCAFS SLC 40 the higher the success rate for the rocket.



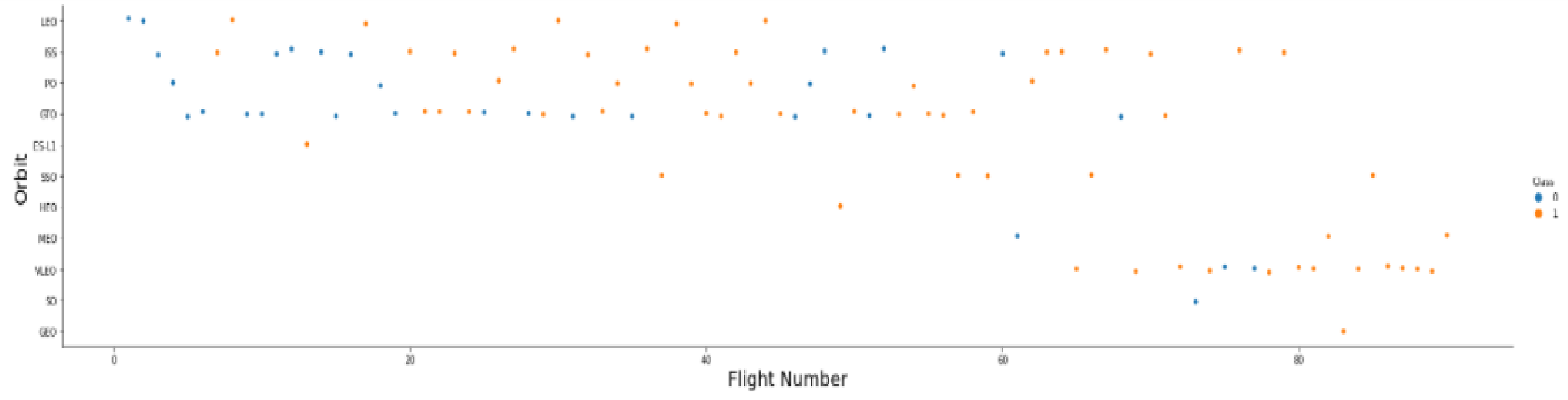
Success Rate vs. Orbit Type

- From the Plot, we can see that ES-L1, GEO, HEO, SSO, VLEO had the most success rate.



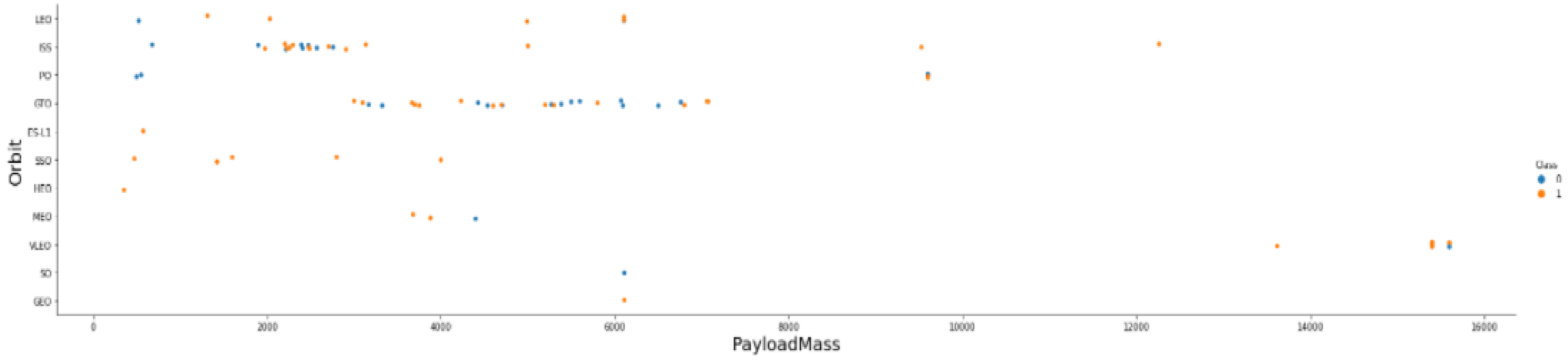
Flight Number vs. Orbit Type

- The plot below shows the Flight Number vs. Orbit type. We observe that in the LEO orbit, success is related to the number of flights whereas in the GTO orbit, there is no relationship between flight number and the orbit.



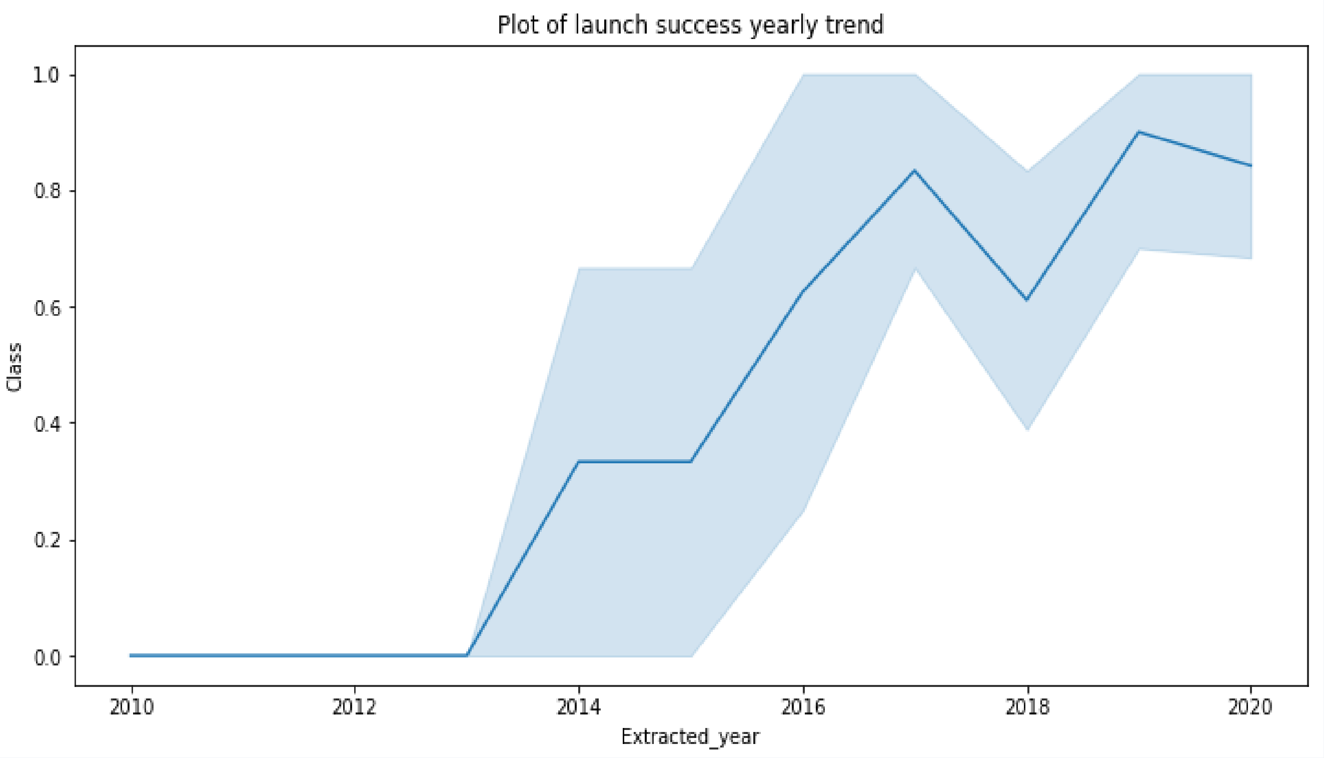
Payload vs. Orbit Type

- We can observe that with heavy payloads, the successful landing are more for PO, LEO and ISS orbits.



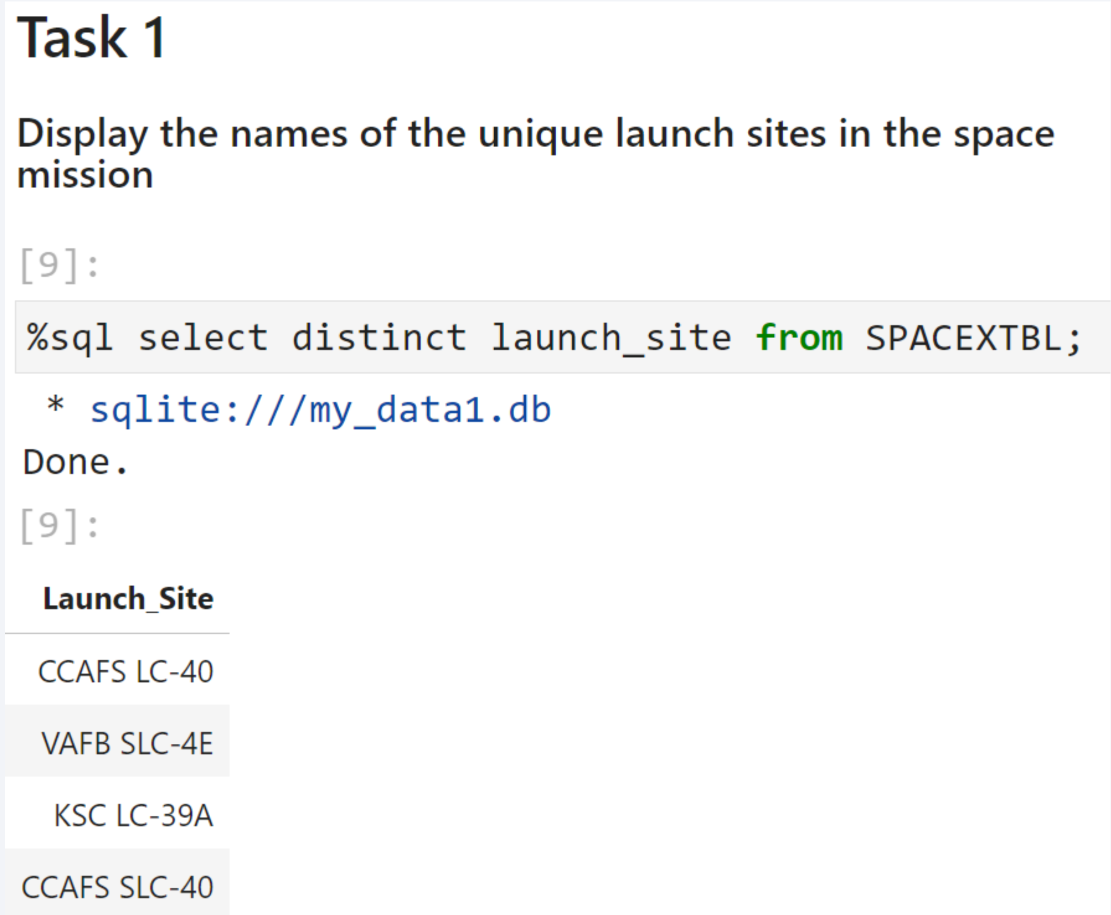
Launch Success Yearly Trend

- From the plot, we can observe that success rate since 2013 kept on increasing till 2020.



All Launch Site Names

- We used the key word **DISTINCT** to show only unique launch sites from the SpaceX data.



Task 1

Display the names of the unique launch sites in the space mission

[9]:

```
%sql select distinct launch_site from SPACEXTBL;
```

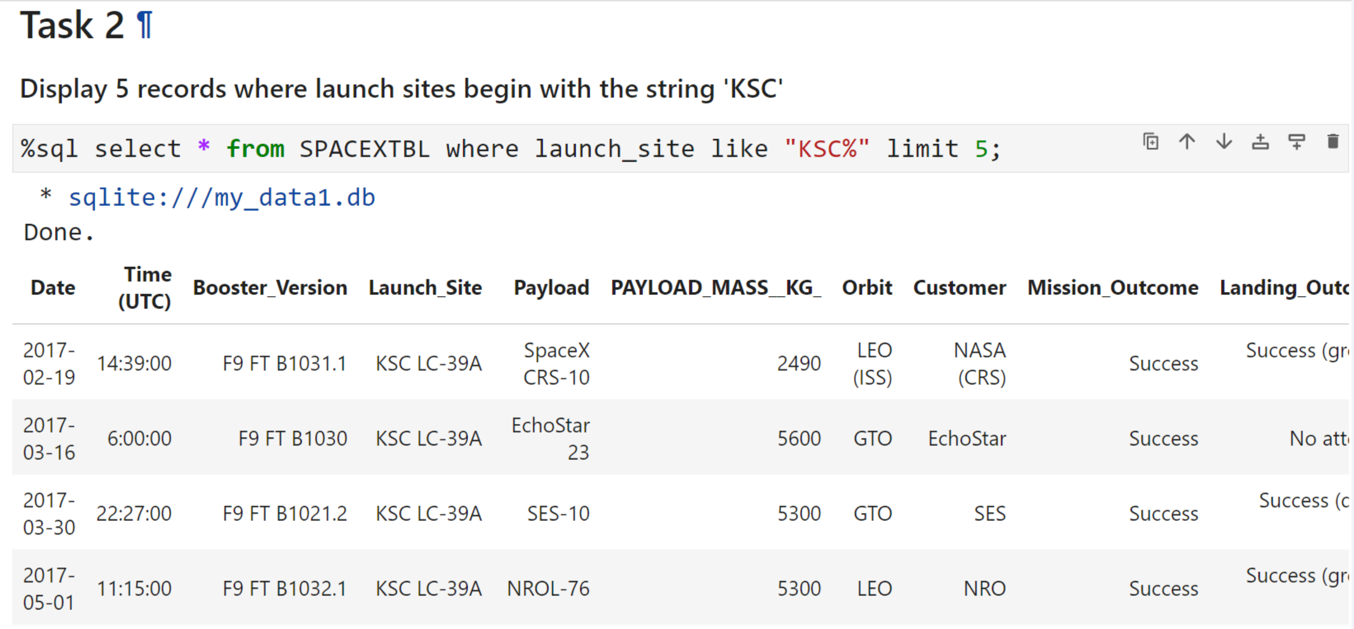
```
 * sqlite:///my_data1.db
Done.
```

[9]:

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

Launch Site Names Begin with 'KSC'

- 5 records where launch sites' names start with `KSC`



Task 2 ¶

Display 5 records where launch sites begin with the string 'KSC'

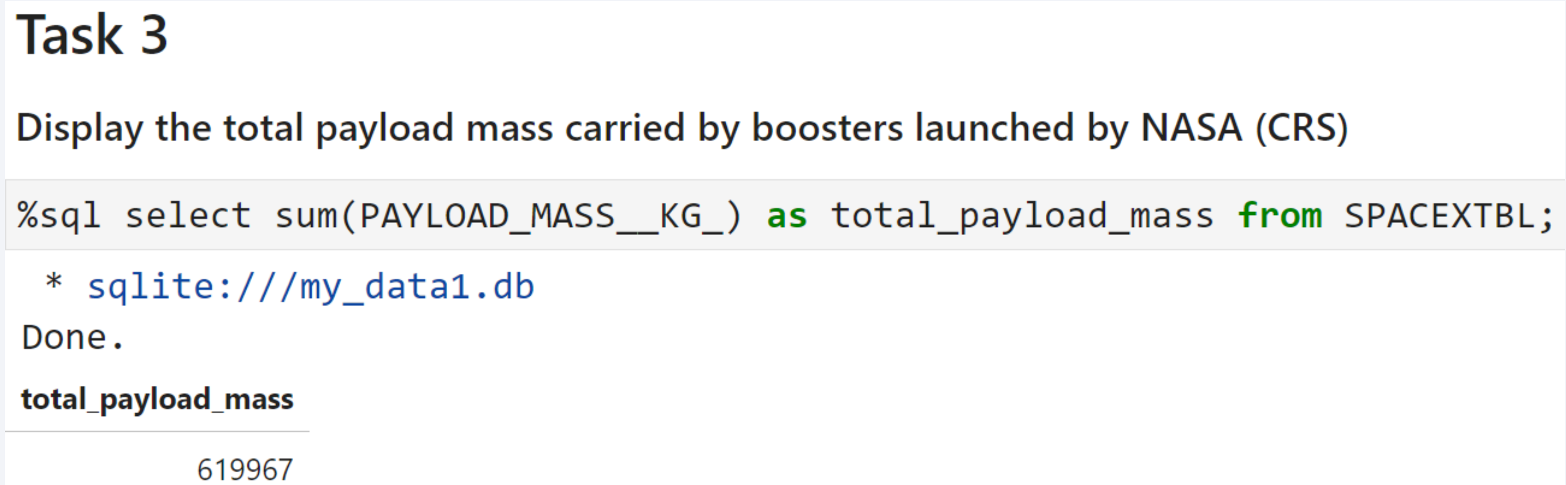
```
%sql select * from SPACEXTBL where launch_site like "KSC%" limit 5;
```

```
 * sqlite:///my_data1.db
Done.
```

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outc
2017-02-19	14:39:00	F9 FT B1031.1	KSC LC-39A	SpaceX CRS-10	2490	LEO (ISS)	NASA (CRS)	Success	Success (gr
2017-03-16	6:00:00	F9 FT B1030	KSC LC-39A	EchoStar 23	5600	GTO	EchoStar	Success	No att
2017-03-30	22:27:00	F9 FT B1021.2	KSC LC-39A	SES-10	5300	GTO	SES	Success	Success (c
2017-05-01	11:15:00	F9 FT B1032.1	KSC LC-39A	NROL-76	5300	LEO	NRO	Success	Success (gr

Total Payload Mass

- The total payload carried by boosters from NASA



Task 3

Display the total payload mass carried by boosters launched by NASA (CRS)

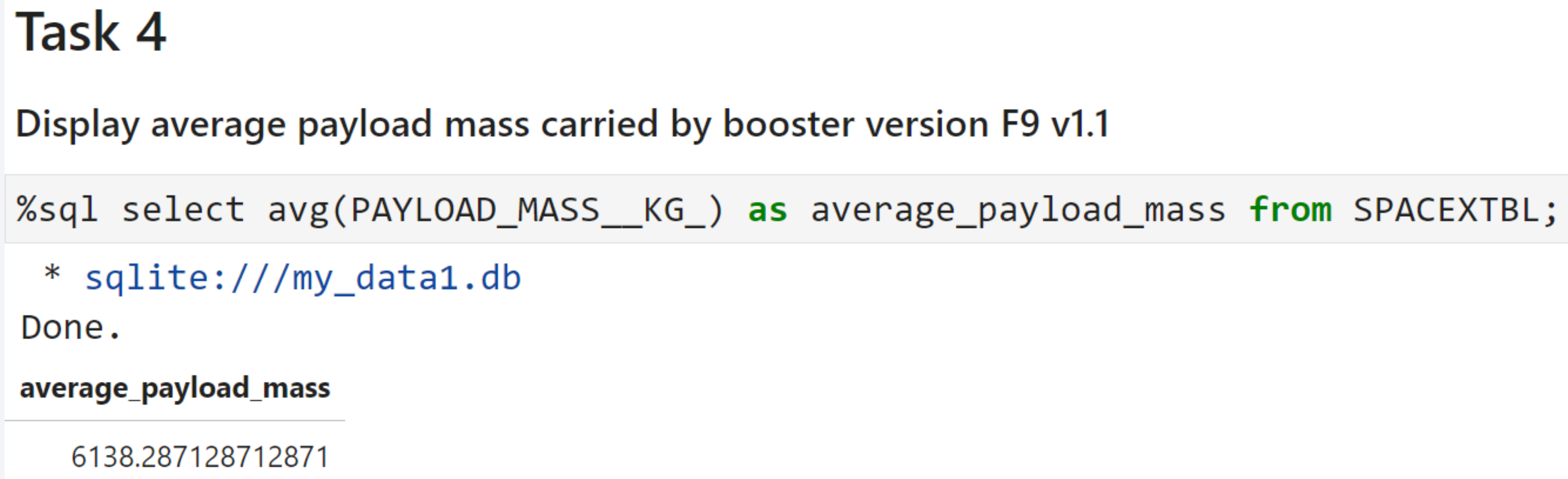
```
%sql select sum(PAYLOAD_MASS__KG_) as total_payload_mass from SPACEXTBL;
```

```
 * sqlite:///my_data1.db
Done.
```

total_payload_mass
619967

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1



Task 4

Display average payload mass carried by booster version F9 v1.1

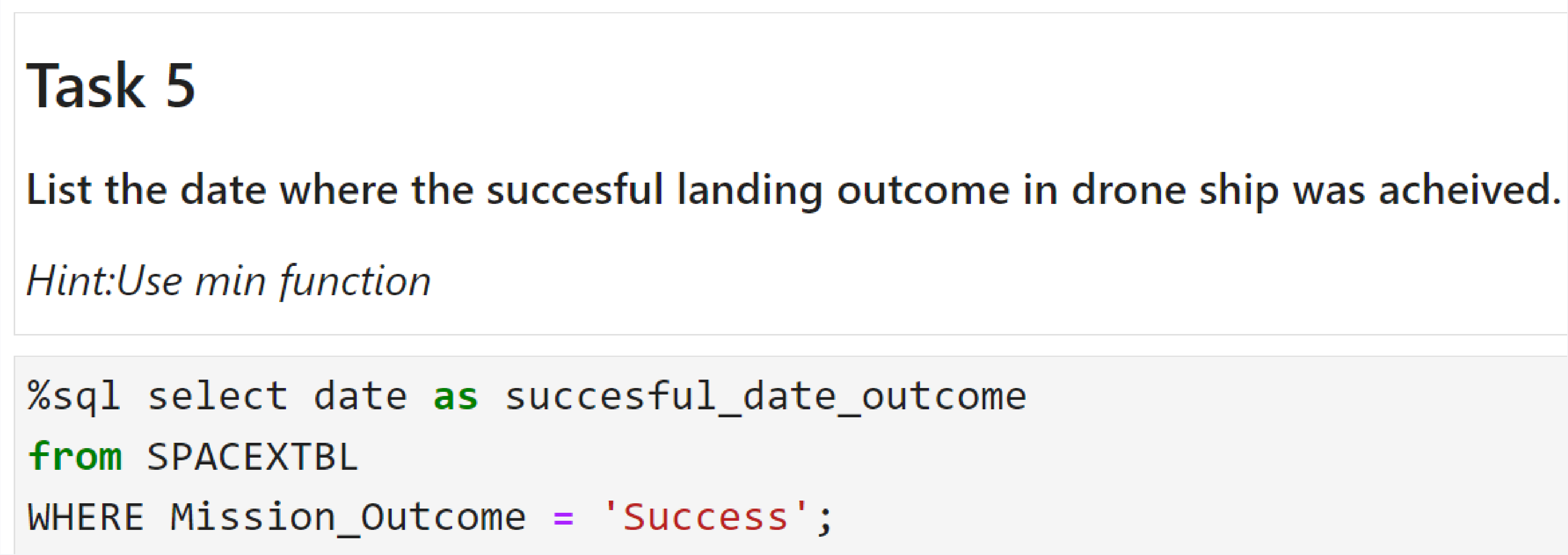
```
%sql select avg(PAYLOAD_MASS__KG_) as average_payload_mass from SPACEXTBL;
```

```
 * sqlite:///my_data1.db
Done.
```

```
average_payload_mass
    6138.287128712871
```

First Successful Ground Landing Date

- We observed that the dates of the first successful landing outcome on ground pad was 22nd December 2015



Task 5

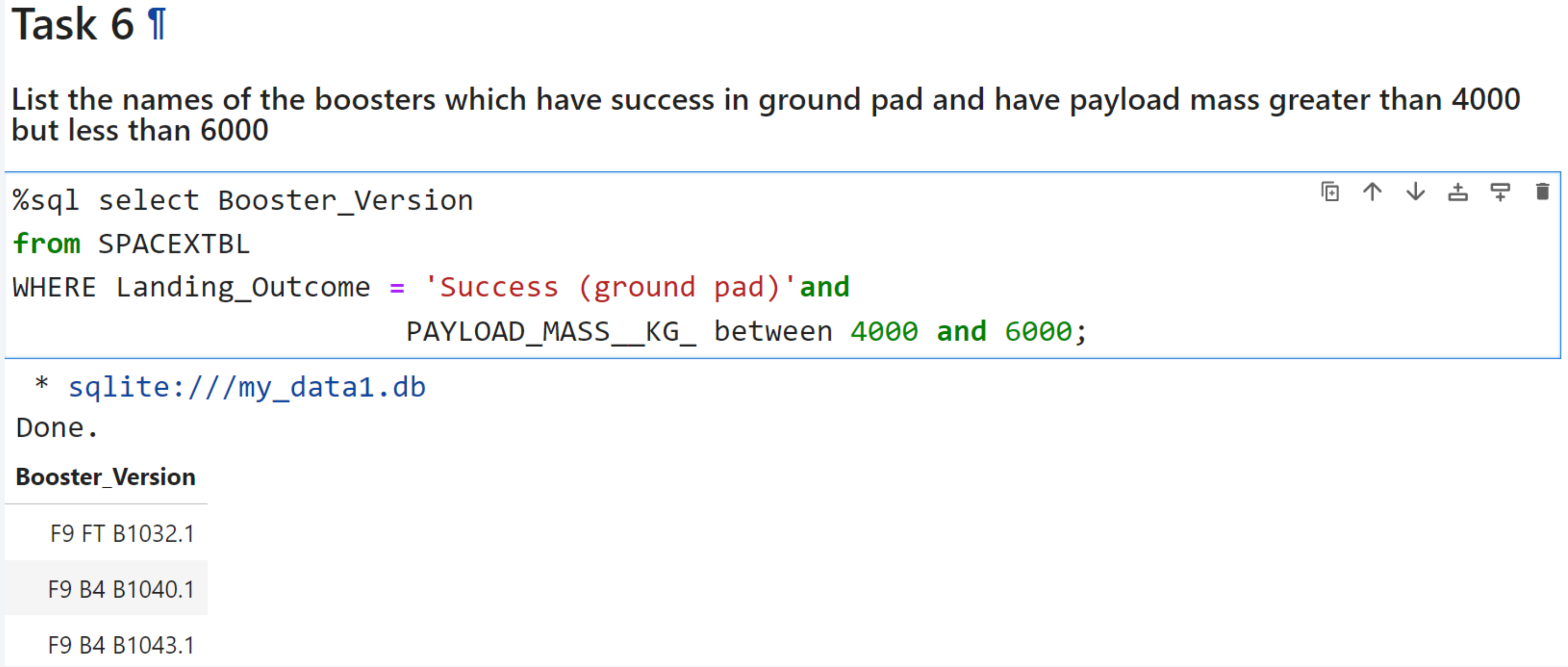
List the date where the succesful landing outcome in drone ship was acheived.

Hint:Use min function

```
%sql select date as succesful_date_outcome
from SPACEXTBL
WHERE Mission_Outcome = 'Success';
```

Successful Drone Ship Landing with Payload between 4000 and 6000

- We used the **WHERE** clause to filter for boosters which have successfully landed on drone ship and applied the **AND** condition to determine successful landing with payload mass greater than 4000 but less than 6000



Task 6 ¶

List the names of the boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000

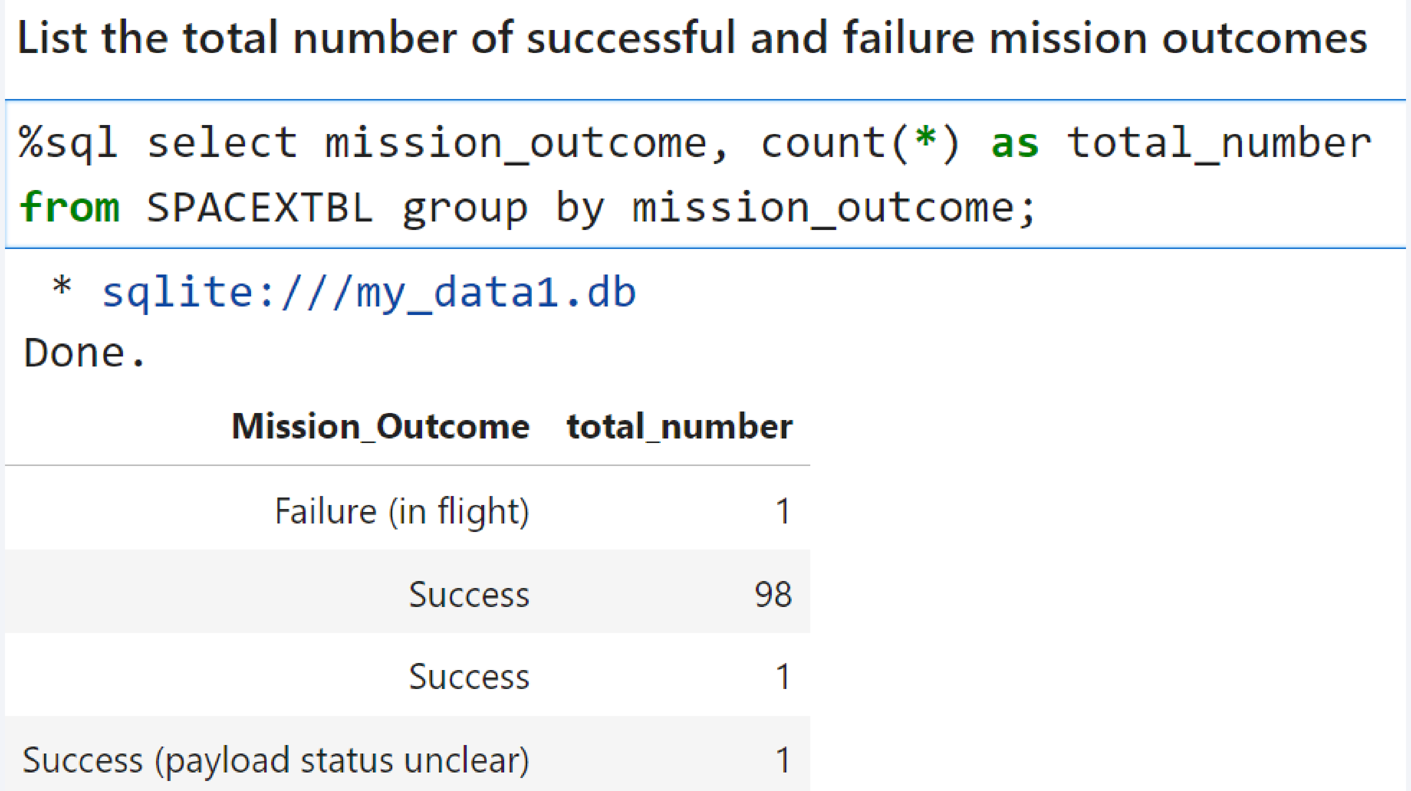
```
%sql select Booster_Version
from SPACEXTBL
WHERE Landing_Outcome = 'Success (ground pad)'and
                      PAYLOAD_MASS__KG_ between 4000 and 6000;
```

```
 * sqlite:///my_data1.db
Done.
```

Booster_Version
F9 FT B1032.1
F9 B4 B1040.1
F9 B4 B1043.1

Total Number of Successful and Failure Mission Outcomes

- We used wildcard like '%' to filter for **WHERE** MissionOutcome was a success or a failure.



List the total number of successful and failure mission outcomes

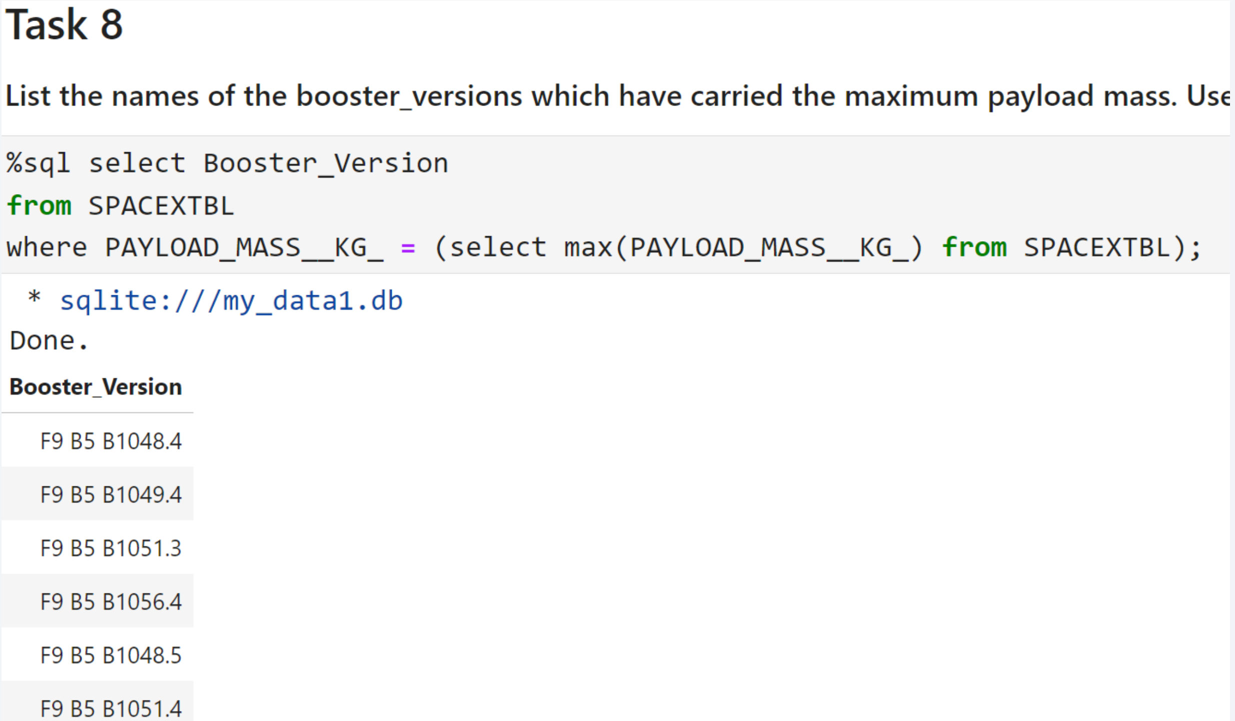
```
%sql select mission_outcome, count(*) as total_number
from SPACEXTBL group by mission_outcome;
```

```
 * sqlite:///my_data1.db
Done.
```

Mission_Outcome	total_number
Failure (in flight)	1
Success	98
Success	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

- We determined the booster that have carried the maximum payload using a subquery in the **WHERE** clause and the **MAX()** function.



Task 8

List the names of the booster_versions which have carried the maximum payload mass. Use

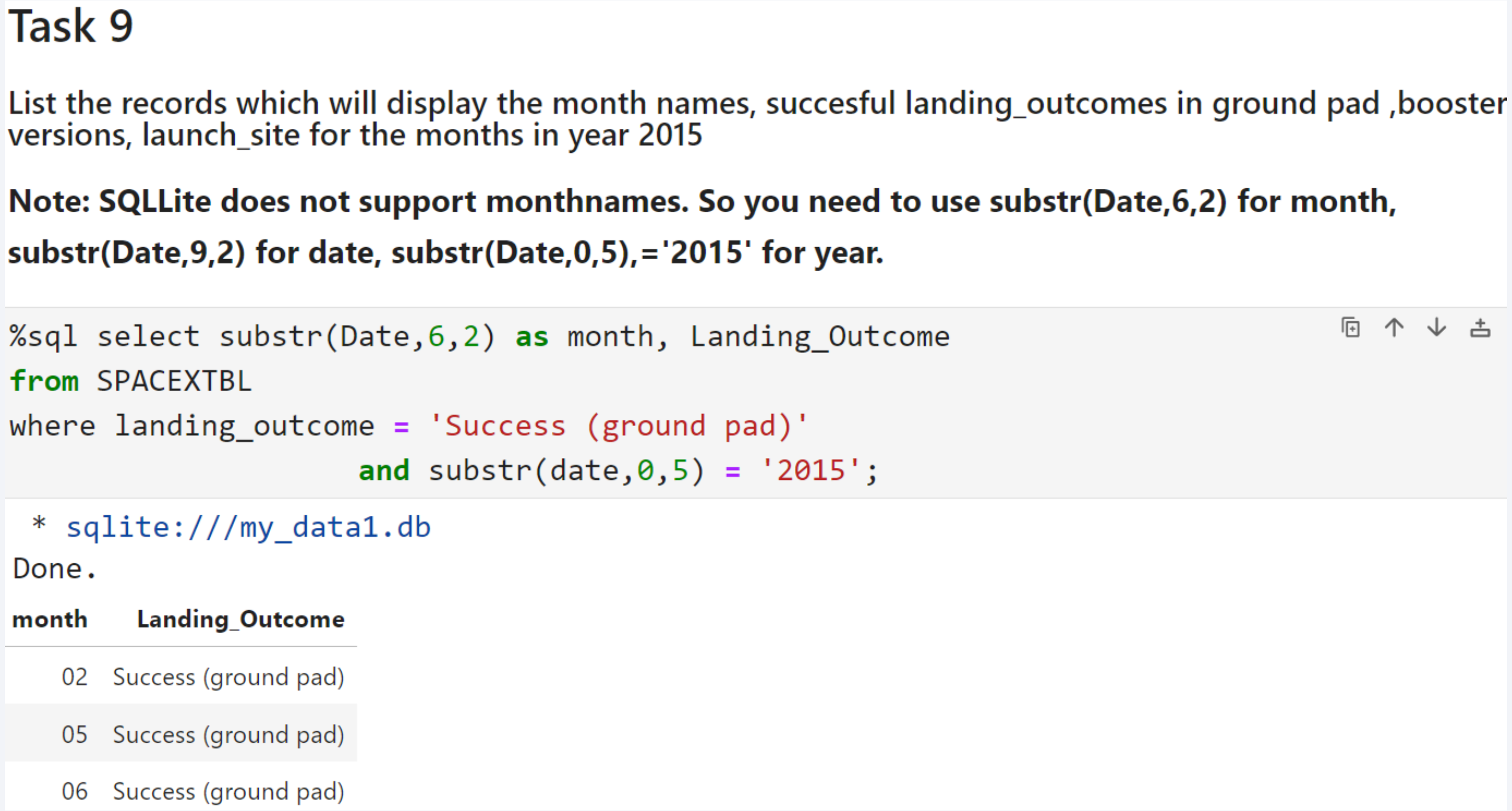
```
%sql select Booster_Version
from SPACEXTBL
where PAYLOAD_MASS__KG_ = (select max(PAYLOAD_MASS__KG_) from SPACEXTBL);
```

```
 * sqlite:///my_data1.db
Done.
```

Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4

2015 Launch Records

- We used a combinations of the **WHERE** clause, **LIKE**, **AND**, and **BETWEEN** conditions to filter for failed landing outcomes in drone ship, their booster versions, and launch site names for year 2015



Task 9

List the records which will display the month names, succesful landing_outcomes in ground pad ,booster versions, launch_site for the months in year 2015

Note: SQLLite does not support monthnames. So you need to use substr(Date,6,2) for month, substr(Date,9,2) for date, substr(Date,0,5),='2015' for year.

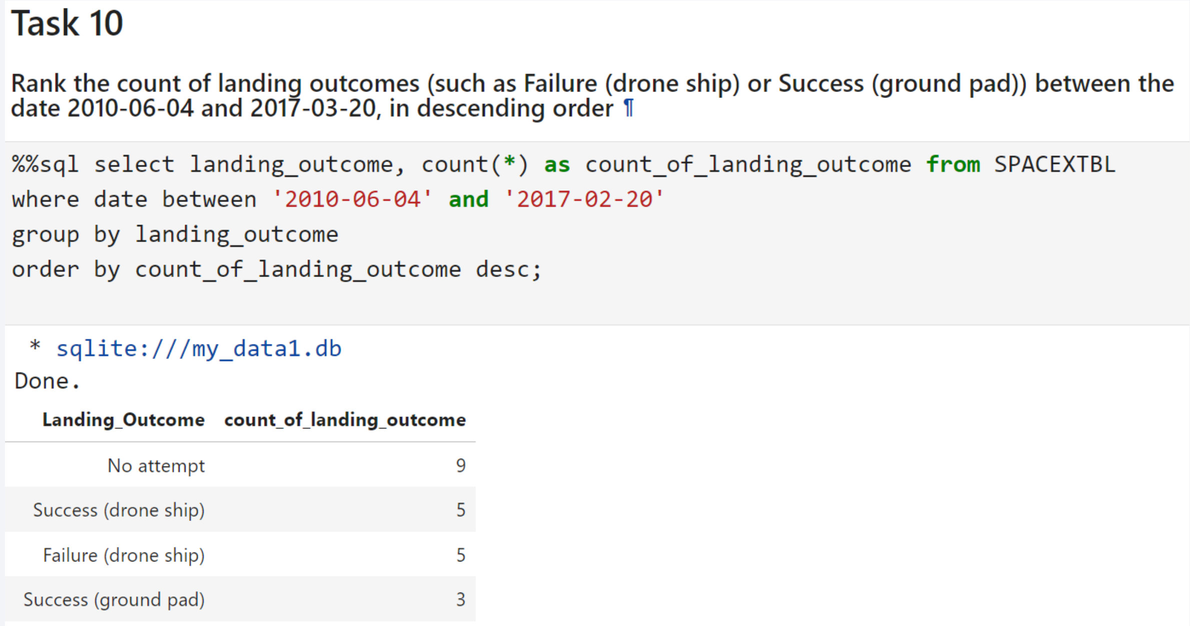
```
%sql select substr(Date,6,2) as month, Landing_Outcome
from SPACEXTBL
where landing_outcome = 'Success (ground pad)'
                    and substr(date,0,5) = '2015';
```

```
 * sqlite:///my_data1.db
Done.
```

month	Landing_Outcome
02	Success (ground pad)
05	Success (ground pad)
06	Success (ground pad)

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- We selected Landing outcomes and the **COUNT** of landing outcomes from the data and used the **WHERE** clause to filter for landing outcomes **BETWEEN** 2010-06-04 to 2010-03-20.
- We applied the **GROUP BY** clause to group the landing outcomes and the **ORDER BY** clause to order the grouped landing outcome in descending order.



Task 10

Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order ¶

```
%%sql select landing_outcome, count(*) as count_of_landing_outcome from SPACEXTBL
where date between '2010-06-04' and '2017-02-20'
group by landing_outcome
order by count_of_landing_outcome desc;
```

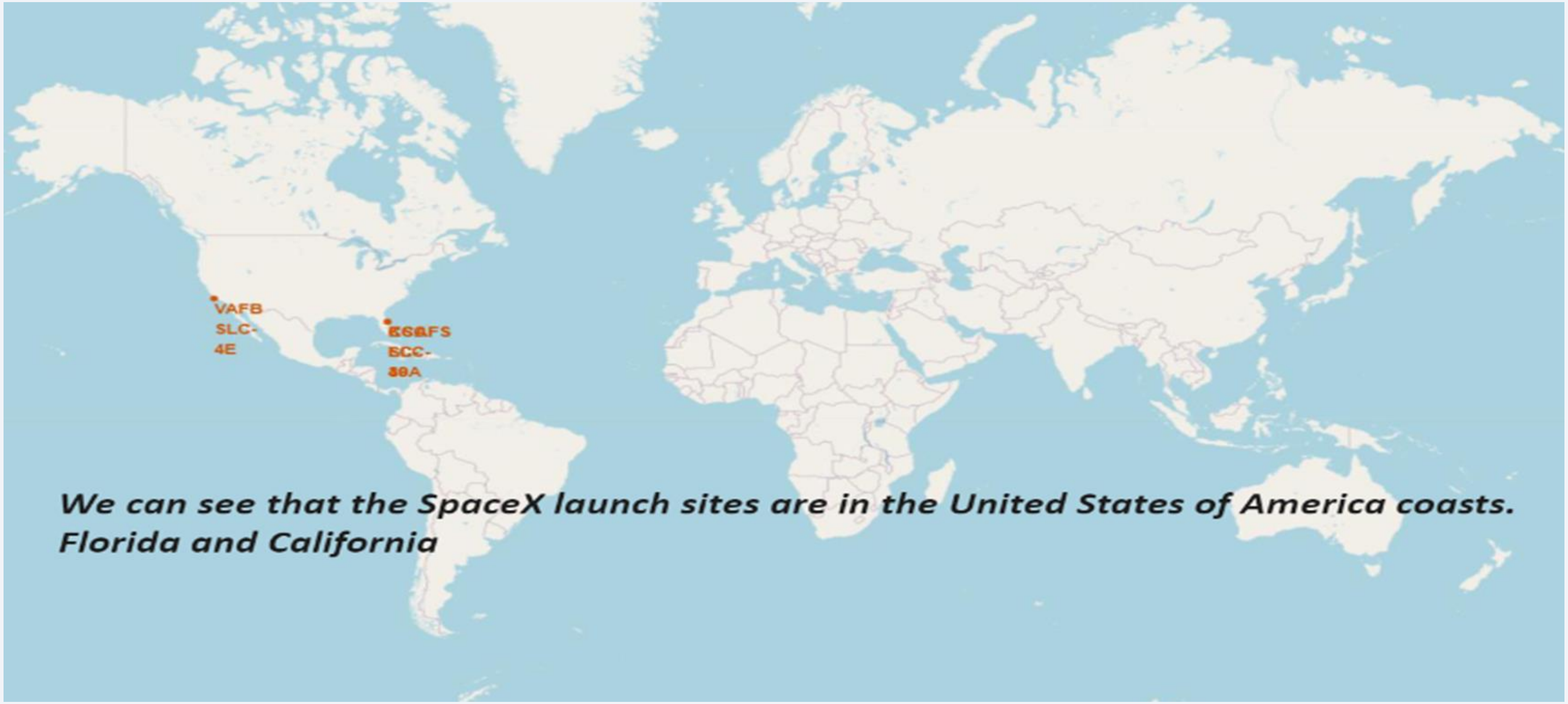
```
 * sqlite:///my_data1.db
Done.
```

Landing_Outcome	count_of_landing_outcome
No attempt	9
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3

Section 3

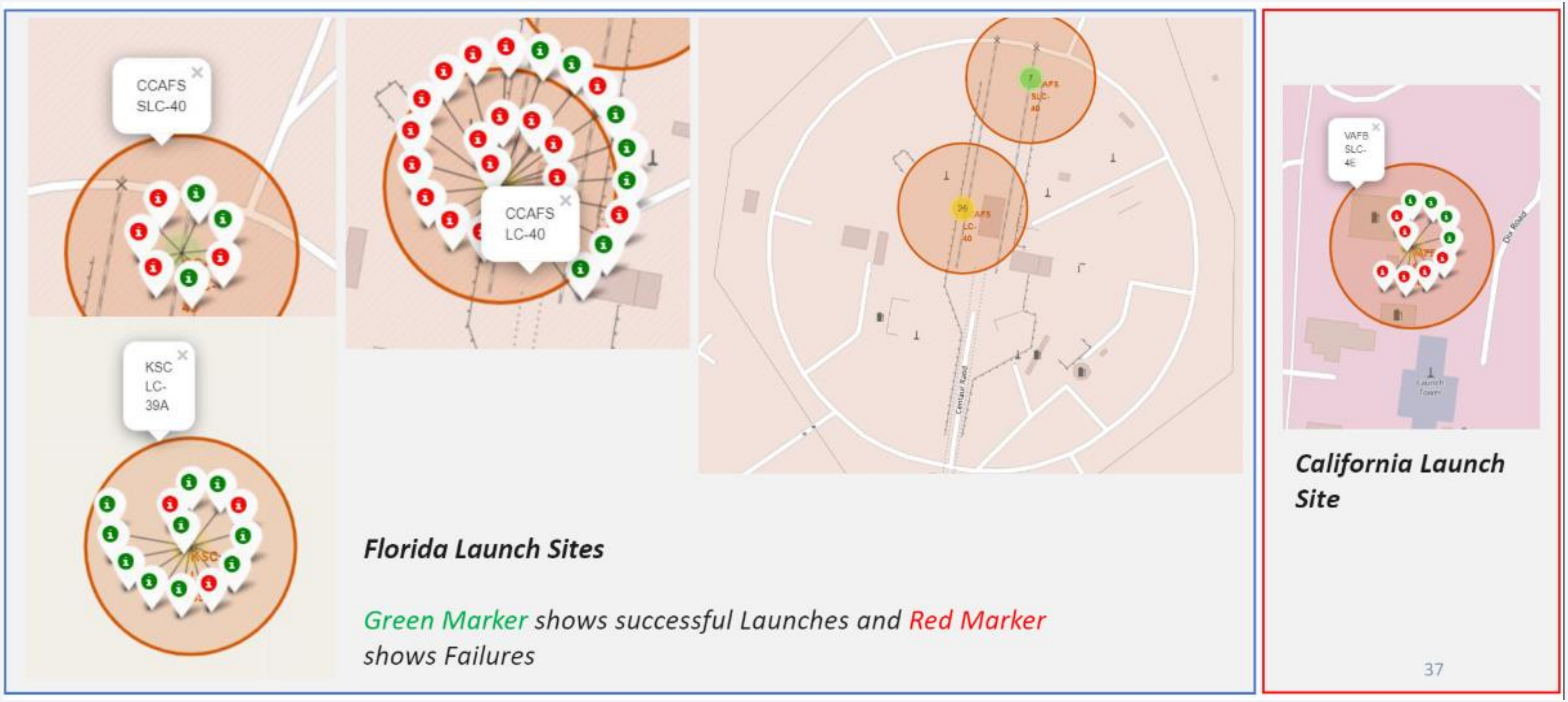
Launch Sites Proximities Analysis

All launch sites global map markers



We can see that the SpaceX launch sites are in the United States of America coasts. Florida and California

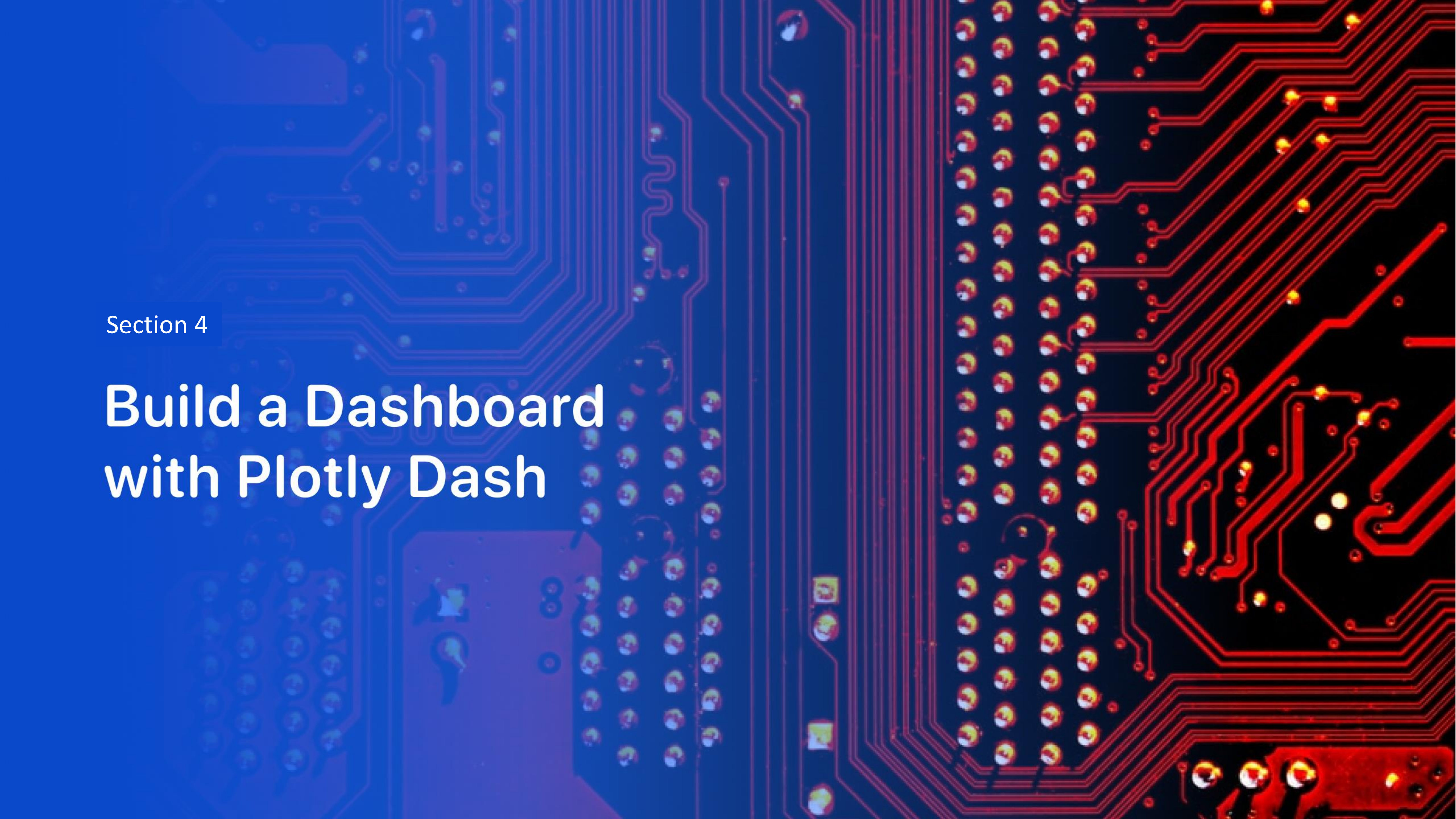
Markers showing launch sites with color labels



Florida Launch Sites

Green Marker shows successful Launches and Red Marker shows Failures

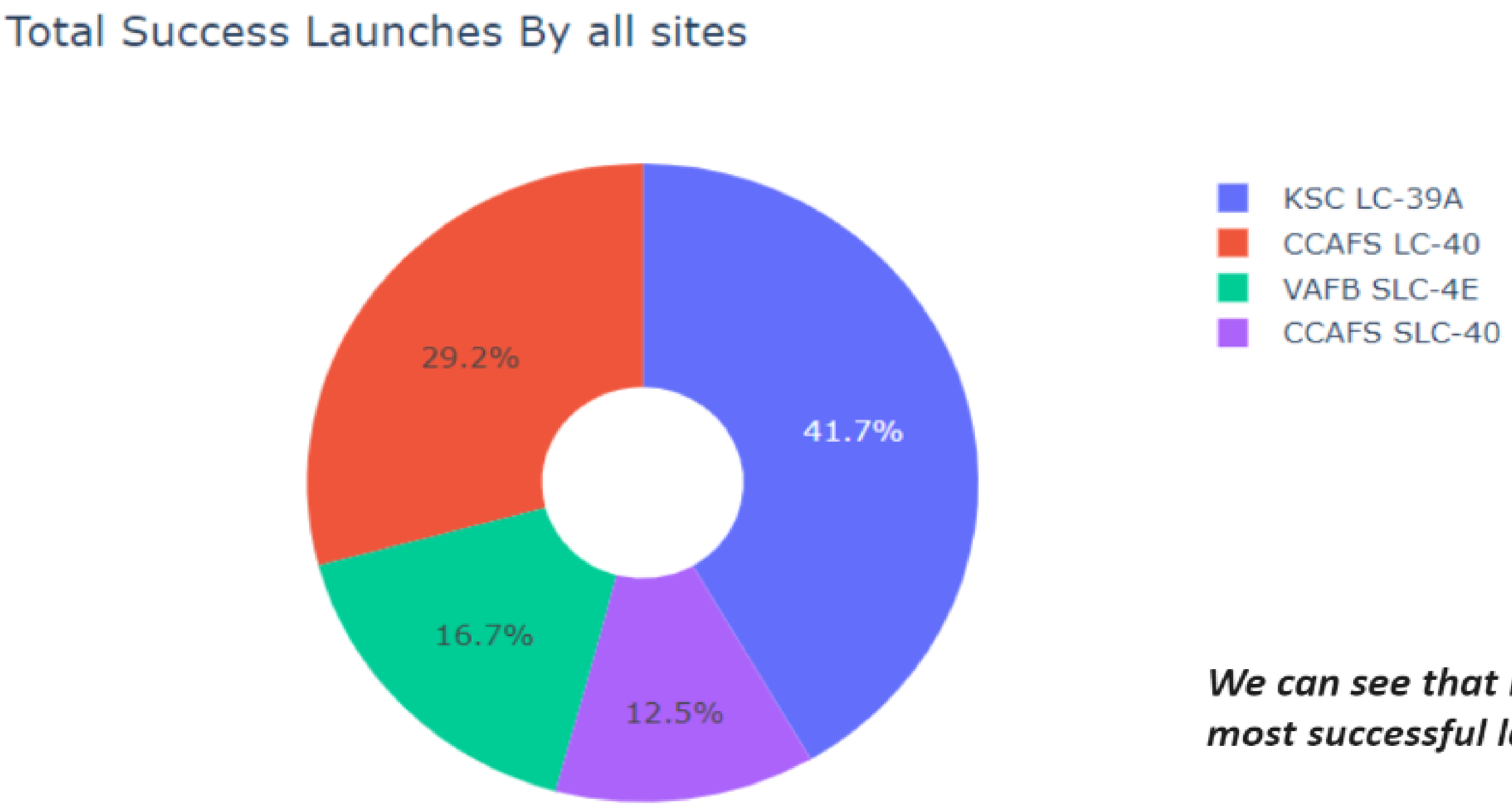
California Launch Site



Section 4

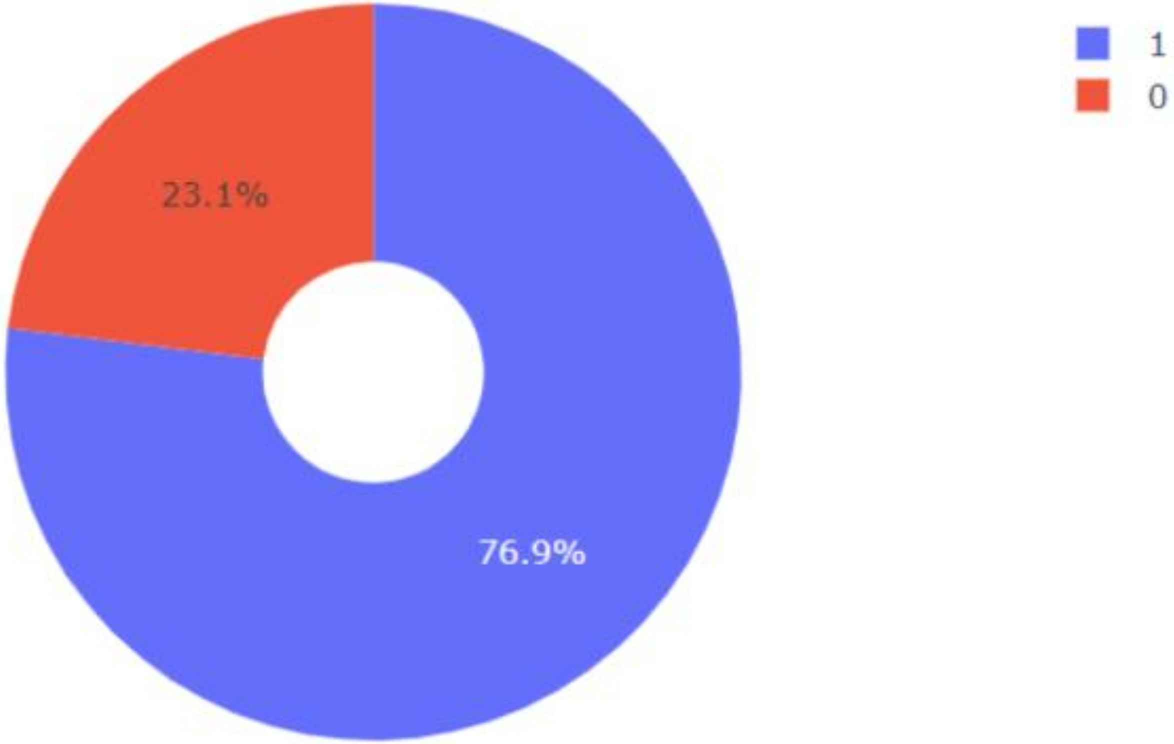
Build a Dashboard with Plotly Dash

Pie chart showing the success percentage achieved by each launch site



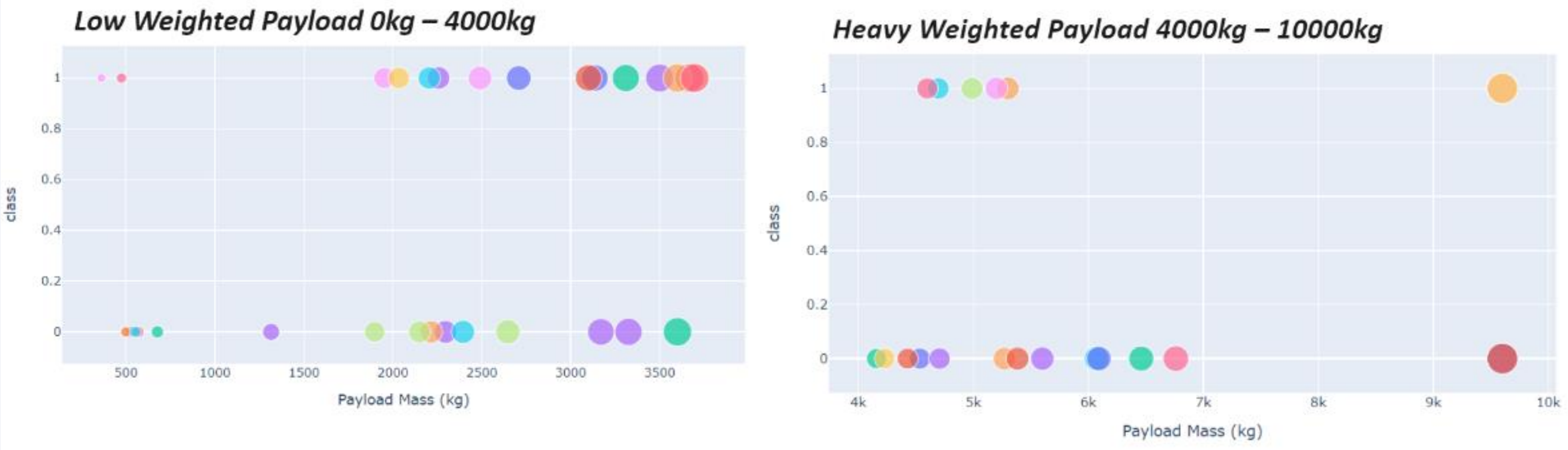
We can see that KSC LC-39A had the most successful launches from all the sites

Pie chart showing the Launch site with the highest launch success ratio



KSC LC-39A achieved a 76.9% success rate while getting a 23.1% failure rate

Scatter plot of Payload vs Launch Outcome for all sites, with different payload selected in the range slider



Low Weighted Payload 0kg – 4000kg

Heavy Weighted Payload 4000kg – 10000kg

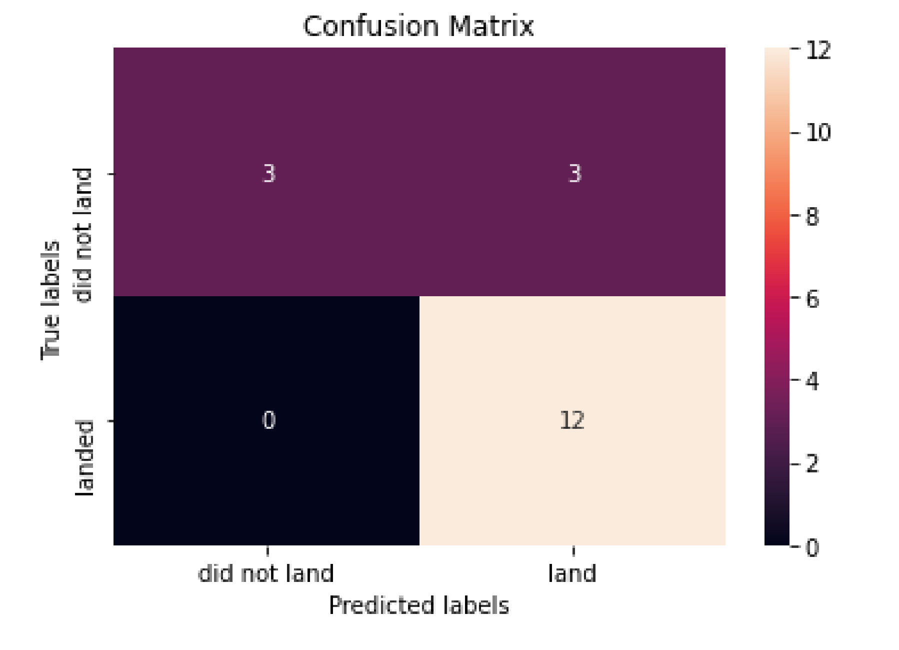
We can see the success rates for low weighted payloads is higher than the heavy weighted payloads

Section 5

Predictive Analysis (Classification)

Confusion Matrix

- The confusion matrix for the decision tree classifier shows that the classifier can distinguish between the different classes. The major problem is the false positives .i.e., unsuccessful landing marked as successful landing by the classifier.



Conclusions

We can conclude that:

- The larger the flight amount at a launch site, the greater the success rate at a launch site.
- Launch success rate started to increase in 2013 till 2020.
- Orbits ES-L1, GEO, HEO, SSO, VLEO had the most success rate.
- KSC LC-39A had the most successful launches of any sites.
- The Decision tree classifier is the best machine learning algorithm for this task.

Thank you!